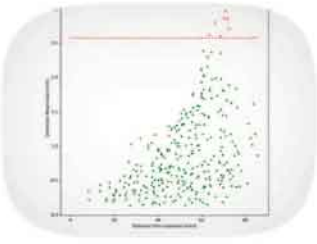
Measure and Evaluate MRgRT 3D Distortion

distortioncheck

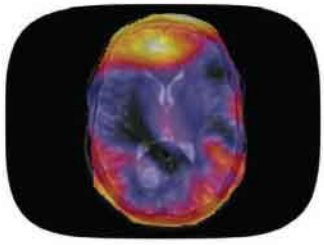
CLOUD SOFTWARE FOR EVALUATION OF IMAGE DISTORTION

SCAN

UPLOAD IMAGES



REVIEW REPORTS & TREND ANALYSIS



EXPORT DICOM OVERLAYS TO TPS

Large Field Grid Phantom
2152 Physical Control Points

- ✔ CIRS proprietary materials simulate distortion due to susceptibility & chemical shifts typical to clinical patient scans
- ✔ Density of physical control points optimized to bring interpolation close to linearity
- ✔ Cloud based solution frees user of operating system and hardware constraints
- ✔ Quickly & automatically analyze complete MR data sets
- ✔ Online deployment facilitates collaboration, easy review and portability of results

Inter-cranial Grid Phantom
335 Physical Control Points

CIRS
Tissue Simulation & Phantom Technology

cirsinc.com
900 Asbury Ave., Norfolk, VA 23513, USA • (800) 617-1177

Complete abdomen and pelvis segmentation using U-net variant architecture

Alexander D. Weston*
Health Sciences Research, Mayo Clinic, 4500 San Pablo Road S, Jacksonville, FL 32250, USA

Panagiotis Korfiatis*, Kenneth A. Philbrick, Gian Marco Conte, Petro Kostandy, Thomas Sakinis,[1] Atefeh Zeinoddini, Arunnit Boonrod,[2] Michael Moynagh, Naoki Takahashi, and Bradley J. Erickson[a)†]
Department of Radiology, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA

(Received 26 February 2020; revised 14 July 2020; accepted for publication 22 July 2020; published xx xxxx xxxx)

Purpose: Organ segmentation of computed tomography (CT) imaging is essential for radiotherapy treatment planning. Treatment planning requires segmentation not only of the affected tissue, but nearby healthy organs-at-risk, which is laborious and time-consuming. We present a fully automated segmentation method based on the three-dimensional (3D) U-Net convolutional neural network (CNN) capable of whole abdomen and pelvis segmentation into 33 unique organ and tissue structures, including tissues that may be overlooked by other automated segmentation approaches such as adipose tissue, skeletal muscle, and connective tissue and vessels. Whole abdomen segmentation is capable of quantifying exposure beyond a handful of organs-at-risk to all tissues within the abdomen.
Methods: Sixty-six (66) CT examinations of 64 individuals were included in the training and validation sets and 18 CT examinations from 16 individuals were included in the test set. All pixels in each examination were segmented by image analysts (with physician correction) and assigned one of 33 labels. Segmentation was performed with a 3D U-Net variant architecture which included residual blocks, and model performance was quantified on 18 test cases. Human interobserver variability (using semiautomated segmentation) was also reported on two scans, and manual interobserver variability of three individuals was reported on one scan. Model performance was also compared to several of the best models reported in the literature for multiple organ segmentation.
Results: The accuracy of the 3D U-Net model ranges from a Dice coefficient of 0.95 in the liver, 0.93 in the kidneys, 0.79 in the pancreas, 0.69 in the adrenals, and 0.51 in the renal arteries. Model accuracy is within 5% of human segmentation in eight of 19 organs and within 10% accuracy in 13 of 19 organs.
Conclusions: The CNN approaches the accuracy of human tracers and on certain complex organs displays more consistent prediction than human tracers. Fully automated deep learning-based segmentation of CT abdomen has the potential to improve both the speed and accuracy of radiotherapy dose prediction for organs-at-risk. [https://doi.org/10.1002/mp.14422]

Key words: abdomen, computed tomography, deep learning, pancreas, gastrointestinal tract, segmentation

1. INTRODUCTION

Radiotherapy treatment planning requires precisely calculating radiation exposure not only of target volumes (visible tumor mass as well as tissue which is likely to contain malignancy) but of nearby organs-at-risk (OARs), that is, organs that will receive substantial dose from the treatment and for which exposure should be minimized. For many abdominal cancers, computed tomography (CT) is the standard of care for radiation treatment planning. Segmentation of OARs (a necessary step for calculating exposure) can be time-consuming and complicated, often taking up to several hours per case. For this reason, OAR segmentation is usually limited only to nearby solid organs which are susceptible to exposure. In the ideal case it may be useful to estimate exposure to all abdominal tissues including adipose, musculoskeletal, and connective tissues.

Deep convolutional neural networks (CNNs) have been utilized to segment increasingly complicated anatomy from medical images.[1,2] Recently, the ability of deep learning models to segment complex and often subtle anatomy has led to its use in tasks involving segmentation of groups of organs[3–6] including whole-region segmentation.[7,8] Such tools would be useful in a variety of medical settings, including radiology, radiation therapy, surgical planning, and 3D printing.[9]

In this work, we propose a 3D end-to-end convolutional neural network capable of complete abdomen and pelvic organ segmentation on CT volumes from individuals with renal-cell carcinoma (RCC). Our goal was whole-body segmentation, that is, to assign each voxel within the body to an appropriate region-of-interest (ROI) corresponding to one of 33 unique organs and tissues, including tissues that may be overlooked by other automated segmentation approaches

such as adipose tissue, skeletal muscle, and connective tissue and vessels.

Access to high-quality annotated data remains the single-biggest limitation to developing deep learning-based approaches. For this report, 84 CT volumes were segmented by ten image analysts and corrected by four radiologists using an open-source annotation tool.[12]

Although manual segmentation is still widely accepted as the gold standard approach, few results on the accuracy of human tracing are found in the literature. In addition to reporting the accuracy of our model, we report metrics for interobserver variability of multiple tracers using semiautomated segmentation on all abdominal organs, and purely manual segmentation on a subset of organs. By including metrics on the accuracy of manual segmentation, we hope to place our results (and the results of similar automated approaches) in the context of the accuracy of existing clinical performance.

Several publications have report simultaneous segmentation of multiple organs. This field has been helped by two publicly available datasets of abdominal CT images and accompanying segmentations provided by *The Cancer Imaging Archive* (TCIA)[13] and *Beyond the Cranial Vault* data challenge sponsored by *Medical Image Computing and Computer Assisted Intervention* (MICCAI).[14] Many models have been developed on these datasets; one of those top performing to-date is presented by Gibson et al. who utilized a fully convolutional 3D U-Net architecture with ROIs downsampled to a resolution of 144 cubic voxels,[4] although many works are competitive, especially Cerrolaza et al.[6] and Hu et al.[15] Roth et al. used a 3D U-Net-based architecture to segment the liver, aorta, portal vein, spleen, stomach, and pancreas, achieving excellent results on a test dataset and also on an external dataset.[5] Currently, the work by Zhou et al. represents the most organs simultaneously segmented by a single model, with 16 structures including large and small intestine and abdominal vasculature including the aorta, inferior vena cava, superior mesenteric artery, celiac artery, and non-specific veins.[3] This work also achieves excellent results by applying three independently trained two-dimensional (2D) U-Nets in the coronal, sagittal, and transverse planes which "vote" on each pixel.

Single-organ segmentation of most of these structures has also been reported in the literature. We have attempted to provide comparative metrics on several single-organ models, but a full discussion is beyond the scope of this work (the work by Sahiner et al. provides a comprehensive review).[9] Pancreas segmentation is reported by Roth et al. demonstrating 3D segmentation in contrast-enhanced CT scans.[1] Wang et al. reported prostate segmentation.[11] The MICCAI challenge for Liver Tumor Segmentation (LiTS) reports excellent results for liver segmentation: (https://competitions.codalab.org/competitions/17094). Lastly, Lessman et al. reported spine segmentation,[16] the work by Zheng et al. demonstrating spinal disc segmentation in MR,[17] and the work by Ibragimov and Xing demonstrating segmentation of the spinal canal.[2]

Our work is novel and significant for several reasons. To the best of our knowledge, at 33 ROIs, it represents the largest number of abdominal tissues segmented simultaneously in the literature. This presents several challenges not reported in other works. For example, we report segmentation of organs that may not appear in every scan (e.g. uterus and prostate). Also, the range in organ volumes (e.g., adrenal glands and liver) attempted by our model is greater than what is reported by other groups. Finally, these examinations are from a patient population with conspicuous pathology, likely representing a greater challenge in segmentation. Our work demonstrates the challenges a deep learning model will encounter in clinical practice, as well as benchmarks for performance (including human interobserver variability metrics) on these challenging datasets.

2. MATERIALS AND METHODS

2.A. Data

Sixty-six (66) CT examinations of 64 individuals were included in the training and validation sets. Eighteen (18) CT examinations from 16 individuals were included in the test set. Fifty-two (52, 65%) individuals were male, 28 (35%) were female. Scans were taken from subjects with renal-cell carcinoma (RCC), but who may have had other diseases in the past (e.g., cholecystitis leading to cholecystectomy). All pixels in each examination were segmented and assigned one of 34 labels: subcutaneous or visceral adipose tissue, muscle, bone, left or right kidney, spleen, pancreas, left or right adrenal gland, left or right renal arteries, left or right renal veins, diaphragm, bladder, uterus or prostate, lumbar discs (all discs between T12 vertebra and S1 vertebra were treated as independent labels), aorta, inferior vena cava, spinal column, stomach, liver, small bowel, large bowel, and gallbladder, and a background class. Figure 1 shows an example of abdominal segmentation. A background class (which represented all external entities including air, table, and blankets) was necessarily included in the model output but is not reported. Finally, soft tissue structures which did not belong to one of these organs were labeled with an additional ROI; this included connective tissue, small vessels of the abdomen, ureters, etc. Not every examination had every tissue class present. Notably, male and female subjects lacked the uterus and prostate, respectively. Hysterectomy and cholecystectomy were also prevalent in this dataset. Because the primary interest was organ segmentation, abnormalities such as tumors, cysts, and surgical implants such as wire or staples were segmented as belonging to their constitutive compartment.

Ground truth segmentations were performed by image analysts and were reviewed and corrected by one of four physicians (P.K., T.S., A.Z., A.B.) with prior experience in abdominal organ segmentation. A total of ten analysts participated in this study. Segmentation was performed using RIL-Contour, an open-source annotation software.[12] Several techniques were used to expedite this process. Soft-tissue, muscle, fat, and bone segmentations were generated using a

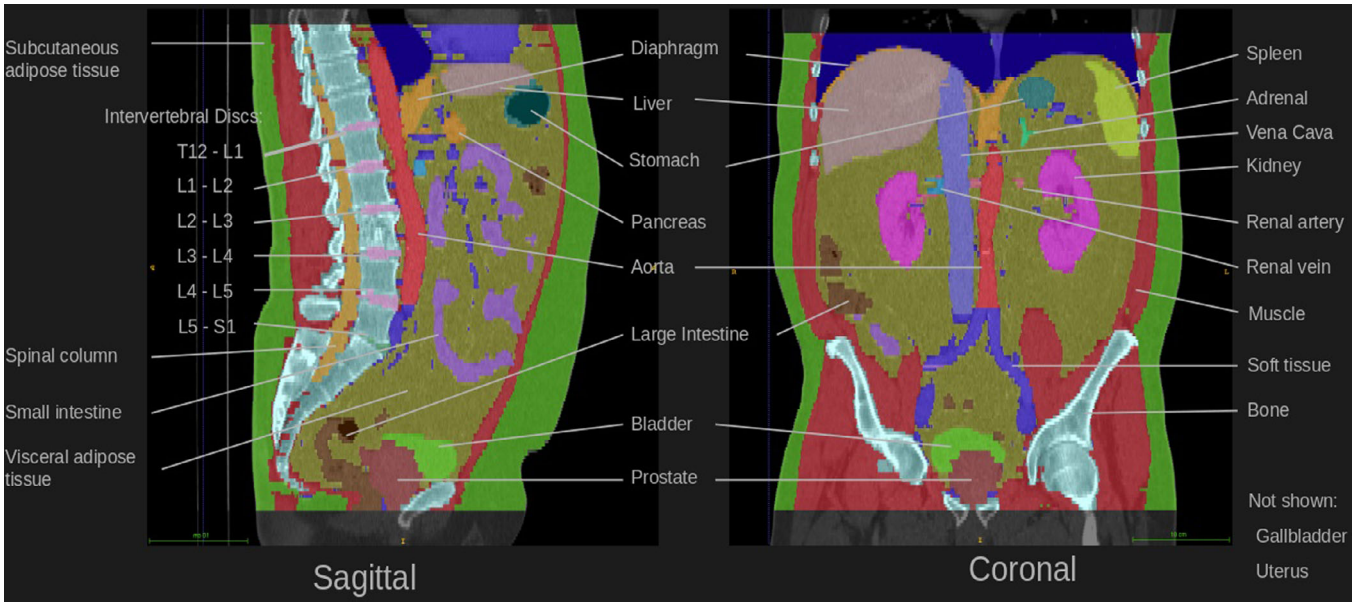


FIG. 1. Sagittal and coronal view of one abdominal computed tomography exam, manually segmented from a machine-generated prior segmentation. For clarity, no distinction is made between left and right structures, although these were differentiated in the training dataset.

previously developed model[18] and were then provided to analysts who corrected existing ROIs and manually segmented novel anatomy. Simple semiautomated tools such as thresholding were also used. Every other slice, or every third slice in cases where the slice thickness was <3 mm, was traced and a full segmentation generated via linear interpolation. In certain cases, multiple analysts worked on the same scan. Any errors or discrepancies in segmentations after interpolation were corrected by a single image analyst.

To measure human interobserver variability, two scans from the test set were segmented by all ten analysts and results combined with the simultaneous truth and performance level estimation (STAPLE) algorithm (the definition of STAPLE algorithm is included in Section 2.C).[25] The performance of each analyst was then compared to this STAPLE-derived ground truth.

Because a machine-generated segmentation was used to aid image analysts, a concern was that this might introduce bias into the dataset. To measure this and to estimate interrater variability of 'free-hand' tracing, several scans were manually segmented by three analysts each without the use of machine-derived segmentations. Because of the effort required, only the liver, spleen, pancreas, kidneys, adrenals, and bladder were segmented.

Other works have achieved volumetric segmentation by annotating multiple slices (but not the whole volume) simultaneously, this is termed the "2.5D" approach.[3] For purposes of this study, full volumes were segmented to preserve the spatial context of organs, important for the segmentation of so many organs. Due to GPU memory constraints, this required resampling scans from native resolution to $256 \times 256 \times 128$ which corresponds approximately to a pixel resolution of $1.5\,\text{mm} \times 1.5\,\text{mm} \times 3\,\text{mm}$. This represents downsampling by a factor of 2 in the anterior–posterior and left–right dimensions but approximately preserves the resolution of the cranial-caudal dimension for scans with a typical slice thickness of 3–5 mm (downsampling did occur for scans with a slice thickness of 1 mm or less).

2.B. Implementation of U-Net architecture

A variation on 3D U-Net architecture was used in this study, as shown in Fig. 2.[10,19] The dimensions of the input of the network were $256 \times 256 \times 128$ cubic voxels. U-Net consisted of two pathways. The first combined downsampling with convolutional layers to encode increasingly abstract representations of the input. The second recombined these representations with shallower features transferred via skip connections to precisely localize the structures of interest.

The activations in the downsampling pathway were computed by so-called context modules. Each context module consisted of a preactivation residual block with two $3 \times 3 \times 3$ convolutional layers with a dropout layer in between. Using residual blocks made the gradient propagation more efficient, ensuring that gradients were capable of backpropagating to lower layers in the network. Context modules are connected by $3 \times 3 \times 3$ convolutions with an input stride of two to reduce the resolution of the feature maps and allow for more features while descending down the aggregation pathway.

The decoding pathway took the low spatial resolution representations and utilized upsampling modules to recreate high-resolution features. Upsampling modules consisted of $2 \times 2 \times 2$ nearest neighbor interpolation followed by $3 \times 3 \times 3$ convolution. The features from the encoding and decoding pathways were concatenated at each level. Work done by Isensee et al. and Kayalibay et al.[20,21] shows that the

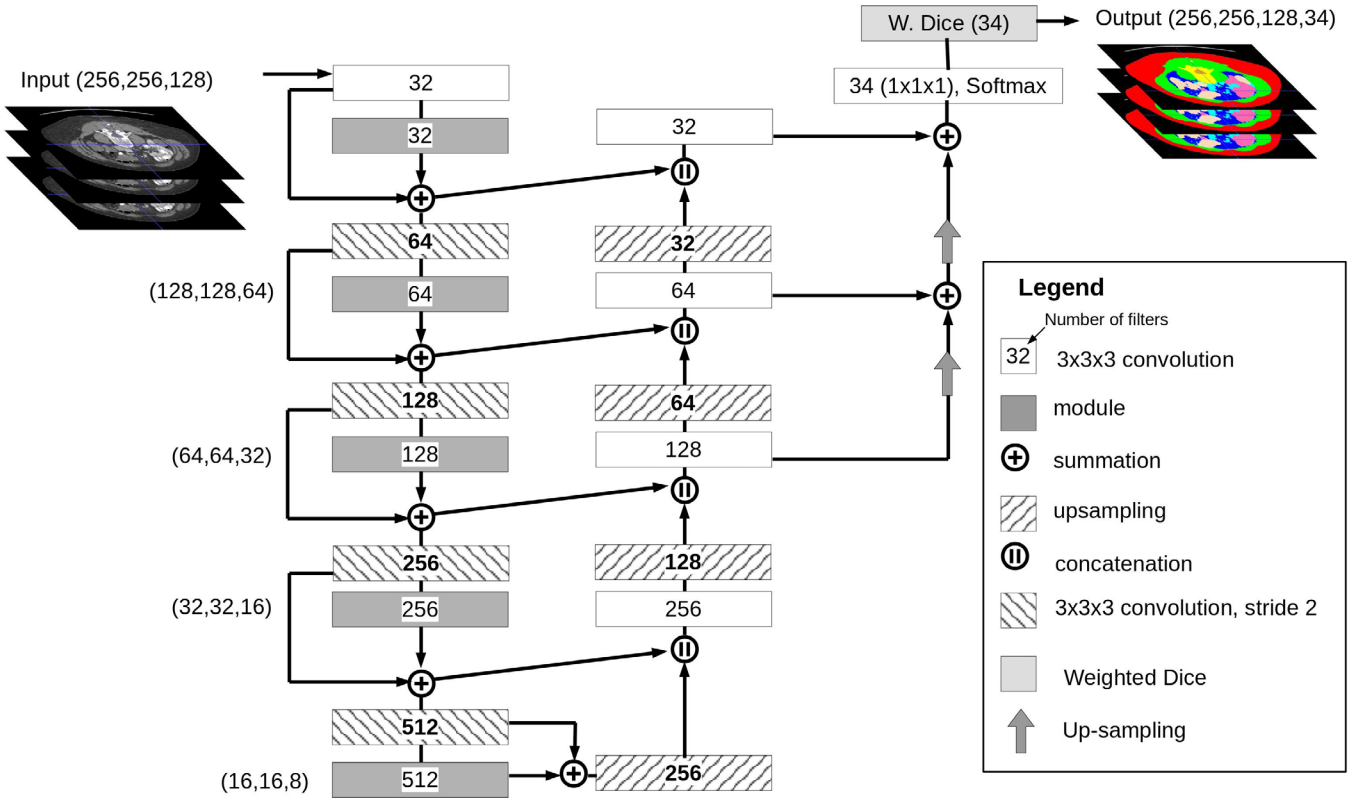


FIG. 2. Schematic of adapted three-dimensional U-Net architecture.

use of multiple layers achieves better localization of the segmentation output.

Throughout the network we used randomized rectified linear unit (RReLU) nonlinearities in which the slope of the negative component of an activation function was randomly assigned within a given range. This type of activation function helped prevent overfitting.[22] Instance normalization[23] was used because it normalized each single image batch independently, i.e., across spatial locations only.

All images were resampled to $256 \times 256 \times 128$ using cubic interpolation. Normalization of the scans was performed with a window width of 1000 HU and center level 100 HU, which was rescaled to the range $[0,1]$. Ground truth segmentations were also downsampled to $256 \times 256 \times 128$ using nearest neighbor interpolation for the purposes of computing accuracy metrics.

Dice score was used as the loss function for all experiments (the definition of the Dice metric is included in Section 2.C). Simple weighting of the individual contributions of the Dice scores was implemented to account for differences in class prevalence. Specifically, the loss contribution of each ROI was weighted in inverse proportion to its prevalence across the training set.

The Dice score on the validation set was monitored and used as a criterion for early termination.

We used the Adam method for optimization with the following parameters: $\beta_1 = 0.9$, $\beta_1 = 0.999$, $e = 1e^{-8}$ An initial learning rate of $1e^{-3}$ was utilized with the "Xavier" initialization.[24]

All networks used in this paper were implemented in Python utilizing Keras 2.2 (open source) with Tensorflow 1.12.0 as backend (open source, Google, Mountain View, CA). All the calculations were performed on systems comprising of a 32 GB card (volta architecture) (NVIDIA, Santa Clara, CA).

The following augmentation strategy was used. Perturbations in the form of rotations, translations and Gaussian white noise were performed on the data. For each of the available cases ten augmented versions were created. Although flipping can be a common augmentation strategy for certain applications, it was not utilized because several organs, for example, liver and pancreas, are not symmetric.

2.C. STAPLE algorithm

A challenge to measuring human accuracy was finding reference segmentations against which to compare human performance. We took the approach of combining multiple human segmentations using the Simultaneous Truth and Performance Level Estimation algorithm (STAPLE).[25]

Consider a set of training data S which consists of a set of N training images $\{I_n\}_{n=1}^{N}$. Each image I_n has been segmented by J humans to generate a set of segmentations $\{Y_{n,j}\}_{j=1}^{J}$. The segmentations $Y_{n,j}$ can be represented by a $D \times R$ matrix of binary values, where D is the voxel dimensions of the scan Y, such that the organ r is either present (1) or absent (0). Ideally, all humans would generate identical

segmentations but the segmentations almost certainly differ. It is also likely that some humans are more accurate than others, that is, one human has the best agreement with a hidden, true segmentation Y^T(which cannot be known).

STAPLE is an iterative Expectation-Maximization algorithm that generates a gold standard Y^{GS} which best approximates Y_T from the individual segmentations $Y_n^{GS} = f\left(\left\{Y_{j,n}\right\}_{j=1}^{J}\right)$. The expectation step for each iteration k is to calculate a set of performance parameters for each analyst $\left\{\theta_j^k \middle| \left(Y_{j,n}, Y^{k-1}\right)\right\}_{j=1}^{J}$. The maximization step is to estimate a new gold standard $Y^{(k+1)} = f\left(\left\{Y_j^{(k)}\right\}_{j=1}^{J} \middle| \theta_j^{(k)}\right)$. The performance metrics θ_j consist of an $R \times R$ matrix describing the probability of an analyst j choosing class r' when the true class is r. Further implementation details are described by Warfield et al.[25]

The algorithm is iterated K times until the rate of change of the true probability Y^K drops below a certain threshold, at which point Y^K is declared Y^{GS}. STAPLE is guaranteed to converge, and generally $K<20$.

2.D. Metrics

Several metrics were used to characterize model performance on the test set. The Dice coefficient between the predicted segmentation Y_P and the reference segmentation Y_R is defined as $\mathrm{Dice}(Y_P, Y_R) = \frac{2*|Y_P \cap Y_R|}{|Y_P|+|Y_R|}$.

The Jaccard coefficient between the predicted segmentation Y_P and the reference segmentation Y_R is defined as $Jaccard(Y_P, Y_R) = \frac{|Y_P \cap Y_R|}{|Y_P \cup Y_R|}$.

The true-positive fraction for a predicted segmentation is defined as $TPF(Y_P, Y_R) = \frac{Y_P \cap Y_R}{|Y_R|}$.

The false-positive fraction for a predicted segmentation Y_P with respect to the reference segmentation Y_R is defined as $FPF(Y_P, Y_R) = \frac{Y_P - |Y_P \cap Y_R|}{|Y_R|}$. For this task, the false-positive fraction is normalized to the voxel count of the reference segmentation because of the high number of negative (e.g., background) voxels.

We also report average volume of the organ predicted by the model, as well as the percent error in volume with respect to the reference segmentation.

3. RESULTS

The average Dice coefficient, Jaccard score, TPF, FPF, predicted volume, and volume error from 18 examinations from the test set are presented in Table I. Volume error represents the absolute value of the difference in volume between the prediction and reference segmentation reported as a percent.

Predictions are compared against semiautomated reference segmentation. For clarity, left and right kidney, left and right

TABLE I. Model performance on 18 test volumes.

Name	Dice	Jaccard	TPF	FPF	Volume (cm^3)	Volume error (%)
Adrenals[a]	0.690 (0.067)	0.531 (0.074)	0.722 (0.102)	0.327 (0.082)	5.0 (2.0)	16 (10)
Aorta	0.884 (0.039)	0.794 (0.060)	0.917 (0.039)	0.142 (0.065)	90.9 (23.5)	8 (8)
Bladder	0.850 (0.081)	0.747 (0.117)	0.857 (0.068)	0.146 (0.125)	163.1 (92.7)	12 (12)
Bone	0.886 (0.012)	0.796 (0.019)	0.892 (0.017)	0.117 (0.023)	1821.8 (550.1)	2 (2)
Colon	0.796 (0.059)	0.666 (0.081)	0.817 (0.080)	0.216 (0.073)	717.4 (256.2)	10 (10)
Diaphragm	0.644 (0.065)	0.478 (0.067)	0.662 (0.066)	0.369 (0.081)	148.5 (54.6)	9 (8)
Gallbladder	0.843 (0.049)	0.732 (0.070)	0.845 (0.036)	0.153 (0.084)	33.5 (13.0)	8 (6)
IV discs[a]	0.644 (0.154)	0.491 (0.143)	0.688 (0.198)	0.360 (0.172)	8.4 (6.3)	28 (24)
IVC	0.830 (0.066)	0.714 (0.089)	0.837 (0.105)	0.166 (0.062)	93.4 (27.5)	11 (12)
Kidneys[a]	0.932 (0.012)	0.874 (0.022)	0.949 (0.013)	0.083 (0.025)	225.5 (50.7)	3 (3)
Liver	0.948 (0.010)	0.902 (0.019)	0.955 (0.012)	0.057 (0.017)	1866.2 (344.6)	2 (1)
Muscle	0.909 (0.008)	0.833 (0.014)	0.920 (0.009)	0.101 (0.020)	8111.8 (2534.0)	3 (2)
Pancreas	0.790 (0.050)	0.656 (0.067)	0.766 (0.081)	0.176 (0.053)	86.6 (19.7)	11 (9)
Prostate	0.776 (0.105)	0.645 (0.132)	0.774 (0.182)	0.191 (0.087)	50.5 (23.7)	21 (19)
Renal arteries[a]	0.507 (0.104)	0.346 (0.093)	0.495 (0.154)	0.436 (0.124)	1.9 (0.9)	34 (25)
Renal veins[a]	0.669 (0.087)	0.509 (0.098)	0.711 (0.122)	0.346 (0.118)	7.2 (4.2)	21 (17)
SAT	0.945 (0.011)	0.896 (0.019)	0.952 (0.009)	0.061 (0.020)	10581.8 (4906.7)	2 (1)
Small intestine	0.814 (0.050)	0.689 (0.067)	0.826 (0.058)	0.195 (0.060)	1023.4 (388.8)	6 (4)
Spinal column	0.848 (0.024)	0.737 (0.037)	0.867 (0.049)	0.166 (0.040)	55.1 (11.8)	8 (5)
Spleen	0.932 (0.021)	0.873 (0.037)	0.941 (0.018)	0.076 (0.030)	273.3 (136.3)	2 (2)
Stomach	0.866 (0.059)	0.768 (0.085)	0.886 (0.042)	0.148 (0.090)	344.1 (247.4)	8 (7)
Uterus	0.634 (0.237)	0.496 (0.250)	0.634 (0.217)	0.305 (0.301)	68.9 (91.7)	40 (30)
VAT	0.892 (0.040)	0.808 (0.062)	0.890 (0.030)	0.103 (0.057)	5484.2 (2506.1)	3 (4)

Results of prediction compared to semiautomated human reference segmentation on 18 test cases. Values represent mean (standard deviation). IVC, inferior vena cava, IV Disks, intervertebral discs; SAT, subcutaneous adipose tissue; VAT, visceral adipose tissue. Metrics reported on downsampled volume.

[a]Values represent the average of multiple structures.

renal arteries, left and right renal veins, adrenal glands, and intervertebral discs are each grouped (individual metrics are provided as Table S1). Our model performed best on the liver with an average Dice score of 0.95 and performed worst on the left and right renal arteries with an average Dice score of 0.51. The difference in organ volume was smallest for the liver with an error of 2% and largest for the uterus with an error of 40%. Volume error tended to be higher for smaller structures. Several scans were missing organs; the uterus was only present in four test cases, the prostate was only present in ten test cases and the gallbladder was only present in twelve test cases. The model failed to predict organs in several cases; the bladder was not segmented in two cases, the L5/S1 disc was not segmented in two cases, and the L4/L5 disc was not segmented in one case. Metrics are reported only for organs that were predicted.

The average accuracy of our model (in terms of the Dice coefficient) compared to human segmentation is reported in Table II. The model was within 5% accuracy compared to human segmentation in eight of the 19 organs reported and within 10% accuracy in 13 of the 19 organs. Semiautomated segmentation values are not reported for the uterus because both scans used to evaluate human segmentation were male.

Semiautomated human segmentation was most accurate on the kidneys with a Dice score of 0.98 and least accurate on the renal arteries with a Dice score of 0.55 compared to STAPLE-derived reference. Manual human segmentation was most accurate on the liver with a Dice score of 0.98 and least accurate on the adrenals with a Dice score of 0.81. The largest discrepancy between model performance and human performance occurred in the adrenal glands (machine: 0.69, human: 0.80), renal veins (model: 0.67, human: 0.79), stomach (model: 0.87, human: 0.97), small intestine (model: 0.71, human: 0.95), and colon (model: 0.80, human: 0.96).

TABLE II. Comparison of model to human performance

Organ	Model	Semiautomated human	Manual human
Kidneys[a]	0.932 (0.012)	0.979 (0.018)	0.975 (0.008)
Bladder	0.850 (0.081)	0.975 (0.011)	0.972 (0.015)
Liver	0.948 (0.010)	0.975 (0.007)	0.980 (0.007)
Spleen	0.932 (0.021)	0.969 (0.011)	0.980 (0.009)
Stomach	0.866 (0.059)	0.965 (0.015)	–
Colon	0.796 (0.059)	0.956 (0.024)	–
Small intestine	0.814 (0.050)	0.950 (0.018)	–
Aorta	0.884 (0.039)	0.924 (0.033)	–
Gallbladder	0.843 (0.049)	0.920 (0.039)	–
Spinal cord	0.848 (0.024)	0.881 (0.030)	–
IVC	0.830 (0.066)	0.869 (0.041)	–
IV Discs[a]	0.644 (0.154)	0.863 (0.103)	–
Prostate	0.776 (0.105)	0.831 (0.076)	–
Adrenals[a]	0.690 (0.067)	0.804 (0.126)	0.810 (0.094)
Renal veins[a]	0.669 (0.087)	0.792 (0.090)	–
Pancreas	0.790 (0.050)	0.754 (0.091)	0.909 (0.045)
Diaphragm	0.644 (0.065)	0.696 (0.068)	–
Renal arteries[a]	0.507 (0.104)	0.549 (0.240)	–
Uterus	0.634 (0.237)	–	–

Average Dice similarity coefficients of model (vs semiautomated human reference), semiautomated human (vs STAPLE-combined semiautomated human reference), and manual human (vs STAPLE-combined manual human reference). Values represent mean (standard deviation).
[a]Values represent the average of multiple structures.

Model accuracy (in terms of the Dice coefficient) is compared to the human accuracy in the box-and-whisker plot in Fig. 3 (note that Fig. 3 plots the median, not the mean Dice score and therefore the ranking does not perfectly match Table II). Although human performance exceeded model performance for all organs, the interquartile range for human performance on the five lowest-scoring organs was greater than the interquartile range for model performance. Model accuracy (in terms of the Dice coefficient) is compared to the best reported results in the literature for multiple organ segmentation in Table III. For comparison, results from several single-organ segmentation models are also reported when the organ is not included in other multi-organ models. In all cases, the results from individual models outperformed our multi-organ U-Net. Our model exceeded the accuracy of the best reported results for a multiple organ prediction model for two of the 19 organ structures reported. Our model was within 5% of the best performance on six organs. Finally, the remaining seven organs were not segmented by other multiple organ models included in the literature. In most cases, model performance was within a few percent with one notable exception: our model demonstrated an average adrenal Dice score of 0.69 compared to a value of 0.36 reported by Zhou et al.[3]

Representative screenshots of five organs are shown in Fig. 4 (screenshots from all 33 organs are included as Figures S1-1, S1-2, S1-3, and S1-4). Model performance is traced in orange; semiautomated human performance is traced in cyan. The Dice score is also displayed.

4. DISCUSSION

In this work, we presented a 3D U-Net trained to perform full abdomen segmentation into 33 organs and tissues. Model results were reported on 18 test scans, and average model performance was comparable to human performance and similar to values reported elsewhere in the literature.

There are challenges to segmenting large scans with many organ structures. Organ volume can vary dramatically, on the test set we measured an average volume of 5 cm^3 for the adrenal gland to an average volume in the liver of 1866 cm^3, and larger for compartments such as visceral or subcutaneous fat. Organs such as the small and large intestine and some blood vessels may be joined by weak connections that do not appear in-plane. Presence of tumor, oral or intravenous contrast, or implantable devices can change the appearance of organs. Finally, in this dataset, not every organ appears in the exam, for example, prostate in female and uterus in male individuals.

There is also an inherent trade-off between segmentation precision and memory usage. CT exams of the abdomen have a minimum spatial resolution of $512 \times 512 \times N$, where N

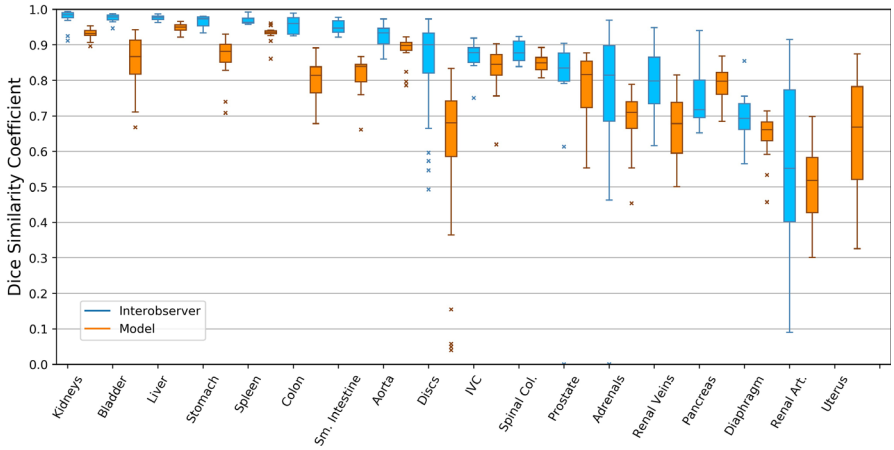


FIG. 3. Average Dice similarity coefficients of model (vs human semiautomated reference) and human (vs STAPLE-combined human semiautomated reference).

TABLE III. Comparison of model to several top-performing models in the literature.

Organ	Model	Zhou[b]	Gibson[c]	Roth[d]	Individual models
Liver	0.95	0.96	0.95	0.97	0.97[e]
Spleen	0.93	0.96	0.95	0.97	–
Kidneys[a]	0.93	0.96	0.93	–	–
Aorta	0.88	0.92	–	–	–
Stomach	0.87	0.93	0.87	0.96	–
Bladder	0.85	–	–	–	0.94[f]
Spinal column	0.85	–	–	–	0.87[g]
Gallbladder	0.84	0.88	0.73	0.84	–
IVC	0.83	0.83	–	–	–
Small intestine	0.81	0.79	0.63[c]	–	–
Large intestine	0.80	0.81	–	–	–
Pancreas	0.79	0.82	0.75	0.87	0.90[h]
Prostate	0.78	–	–	–	0.89[f]
Adrenals[a]	0.69	0.36	–	–	–
Renal veins[a]	0.67	0.70[b]	–	0.79	–
IV discs[a]	0.65	–	–	–	0.92[i]

[a]Values represent the average of multiple structures.
[b]Zhou et al. [3], report metrics for nonspecific veins which we have compared to renal veins
[c]Gibson et al. [4] reports metrics for duodenum which we have compared to small intestine.
[d]Roth et al. [5].
[e]Results taken from the best-performing model to-date on the LiTS Challenge, *mastermind*.
[f]Wang et al. [10].
[g]Ibramigov et al. [2].
[h]Roth et al. [5].
[i]Zheng et al. [16].

represents the number of slices and ranges from approximately 100 slices to over 500 slices. A single CT exam therefore typically has a file size ranging from 100 to 300 MB, which is a challenge both in terms of GPU memory usage and for designing a neural network architecture that can fully capture the feature space. Our strategy was to downsample the image to a lower resolution, which preserves spatial context at the cost of high-resolution features,[10] other strategies include segmenting small patches of the original image which can preserve spatial resolution at the cost of spatial context as well as a reduced receptive field of the network[2,11] and applying 2D models on a slice-by-slice basis at the cost of 3D spatial context.[3]

In general, the true-positive error and false-positive error for most organs were identical, indicating that the model neither consistently over-segments nor under-segments certain organs, a concern when there is substantial class-imbalance (which in this case refers to the large differences in organ volumes). Two exceptions were the pancreas (TPF = 0.766, FPF = 0.177, indicating the model is under-segmenting) and renal arteries (TPF = 0.495, FPF = 0.437, indicating the model is over-segmenting). Our results still lagged behind previously reported single-organ models. This may be due to the fact that other models were optimized for specific organ anatomy, whereas our dataset had significant variability. The model was unaffected by presence of tumor burden, segmentation of the kidneys (the affected organ in this study) was third-most accurate in this study with a Dice coefficient of 0.93.

In this study, several test examples had organs which were absent. Eight out of 18 cases were missing at least one organ. Six cases were missing the gallbladder due to surgical resection. Two scans were truncated mid-abdomen: in these cases, the L5/S1 disc, prostate or uterus, and bladder were not visible in either scans. Finally, in two cases our model predicted prostate when the ground truth did not contain prostate. In both cases the scan did not extend to the pubic symphysis but instead truncated just inferior to the bladder. On further inspection, we determined that a portion of the prostate was indeed visible in the scan and that the prediction was correct and the ground truth was not. Although model accuracy on organs such as the prostate, gallbladder, and uterus is strong, we believe better performance may be achieved by training a second, preliminary neural network to identify whether the organ is present in the scan.

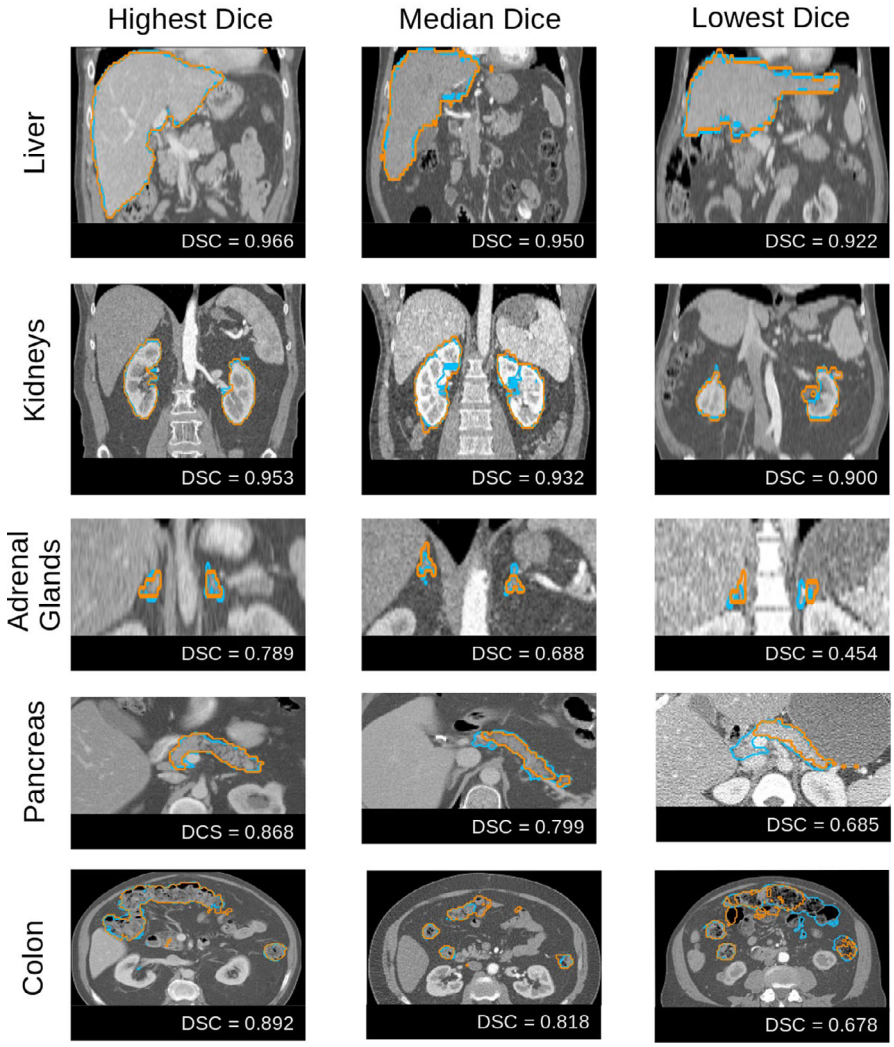


FIG. 4. Representative segmentations of five organs taken from the test set. Model performance is in orange; human semiautomated reference in cyan. The Dice score (DSC) is also displayed.

For several organs in the digestive tract, manual segmentation exceeded model performance by a substantial margin. Gastrointestinal organs are often difficult to segment due to their variability in location and shape which often includes weak, out-of-plane connections, as well as contrast and texture of tract contents and occasional presence of oral contrast. As has been discussed elsewhere, segmentation of the digestive tract has rarely been reported and model accuracy is often poor.[4]

Metrics for model performance and human performance were highest for large organs, for example, liver and small and large intestine, and worse in small organs, for example, adrenals, renal arteries and veins, and diaphragm. This is partially the effect of using the Dice similarity coefficient to assess accuracy, as Dice tends to be high in large, solid organs due to the high volume of overlap, even if accuracy at the edges is low. Segmentation snapshots reveal mistakes in the large organs in both manual and model segmentation.

The error in volume between the prediction and reference varied between 2% and 40%, but generally was lower than the error in Dice score. This is relevant because segmentation in radiology is often performed for volumetry, for example, in the case of evaluating liver volume to assess liver transplant, tracking of kidney size in polycystic kidney disease, and evaluating fatty pancreas disease.[26] In these cases, the error in volume would be 2%, 3% and 11%, respectively. Furthermore, interobserver variability as measured by the interquartile range of the Dice scores between human tracers often exceeded our model on the smallest organs, indicating that the model may be more consistent than manual segmentation.

Few publications report human performance metrics for medical image segmentation which can make it difficult to place model performance in context. We reported the average accuracy (in terms of Dice similarity coefficient) for semiautomated human segmentation (who were provided a "prior" machine-generated segmentation) and for manual human segmentation (without the use of semiautomated tools) on a subset of organs. Model performance, both our results and what is reported elsewhere in the literature, still lagged behind

human performance for most organs. This is in contrast with classification tasks such as the ImageNet challenge[24] where the performance of a deep-neural network has exceeded human performance for some time. This may reflect the limited training set for medical segmentation, vs the millions of examples in ImageNet challenge.

5. LIMITATIONS

Data were taken from individuals with renal-cell carcinoma, many of whom had other findings such as cholecystectomy or prostatectomy. Renal tumors were visible in almost all cases and are observable in Fig. 3. The population of our dataset tended to be older, male, and in frail health. Many scans contained surgical implants or other abnormal anatomy; these features were segmented as belonging to their constitutive compartment. This was an intentional decision intended to reduce the variability in segmentation; even for experts it is often unclear how to classify various medical implants. Furthermore, several scans excluded pelvic organs. For these reasons the low numbers of cases with intact uterus and prostate remained a limitation.

Due to memory constraints, both scans and segmentations were downsampled by a factor of 2 in the coronal and sagittal dimensions. This may introduce partial volume error in the smallest organs such as adrenals. Axial dimension (slice thickness) was resampled to 128 slices, which may have resulted in either slight upsampling or downsampling, but in general preserved slice thickness.

An attempt was made to present these results in context by reporting on the Dice scores and organ volume of manual segmentation. In this study, all scans were acquired on Siemens brand scanners with similar acquisition parameters. Our work and the work of others is generally reported only on test datasets which are derived from internal data, limiting direct comparison. Contextual information such as organ connectivity can be poorly captured by the Dice score.

Finally, human tracers were provided with a prior, machine-generated segmentation which may have reduced interobserver variability and reinforced overfitting behavior during the training process. The average performance of segmentation was slightly lower for an analyst working from a machine segmentation, however, the difference in most cases was $<1\%$. This method of augmenting human performance with machine learning accelerates segmentation so dramatically that we believe it will become increasingly more common both in research and in clinical practice. It may be useful to more carefully characterize the common failings of deep learning models to fully understand what types of corrections may be needed in augmented human intelligence workflows.

6. CONCLUSIONS

The proposed U-Net architecture can segment the abdomen and pelvis into 33 unique organ and tissue structures. This represents the most structures simultaneously segmented from abdominal CT to-date. The model was within 5% accuracy compared to human segmentation for eight of the 19 organs reported and within 10% accuracy for 13 of the 19 organs. Several challenges of multiple organ segmentation include complex anatomy and the difficulty of addressing volume imbalance. Automated segmentation has the potential to automate routine clinical calculations as well as several clinical tasks including surgical guidance and 3D printing, and radiotherapy planning.

ACKNOWLEDGMENTS

Dr. Weston is supported by Mayo Clinic Graduate School of Biomedical Sciences. Funding support was also provided by the Department of Radiology, Mayo Clinic. The authors would like to thank Zeynettin Akkus, Jason Cai, Scott Squires, Bill Ryan, Dan Hertzfeldt, Rachel Marks, Clare Buntrock, Isabel Bazley, Cailin Austin, Nicholas DeBlois, Lucas Betts, Megan Berry, Margaret Cantlon, Angela Weiler, Bailey Ullom, and Lindsey Anding.

CONFLICT OF INTERESTS

The authors have no conflict to disclose.

*co-first authors.
†senior author.
[1]Present address: Department of Radiology, University of Oslo, Oslo Norway
[2]Present address: Khon Kaen University, Khon Kaen Thailand
[a)]Author to whom correspondence should be addressed. Electronic mail: bje@mayo.edu.

REFERENCES

1. Roth H, Oda M, Shimizu N, et al. Towards dense volumetric pancreas segmentation in CT using 3D fully convolutional networks. Paper presented at: Medical Imaging 2018: Image Processing; 2018.
2. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys*. 2017;44:547–557.
3. Zhou Y, Wang Y, Tang P, et al. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. Paper presented at: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV); 2019.
4. Gibson E, Giganti F, Hu Y, et al. Automatic multi-organ segmentation on abdominal CT with dense v-networks. *IEEE Trans Med Imaging*. 2018;37:1822–1834.
5. Roth HR, Oda H, Hayashi Y, et al. Hierarchical 3D fully convolutional networks for multi-organ segmentation. arXiv preprint arXiv:170406382; 2017.
6. Cerrolaza JJ, Picazo ML, Humbert L, et al. Computational anatomy for multi-organ analysis in medical imaging: a review. *Med Image Analysis*. 2019;56:44–67.
7. Zhu W, Huang Y, Zeng L, et al. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys*. 2019;46:576–589.
8. Chen Y, Ruan D, Xiao J, et al.Fully Automated Multi-Organ Segmentation in Abdominal Magnetic Resonance Imaging with Deep Neural Networks. arXiv preprint arXiv:191211000; 2019.
9. Sahiner B, Pezeshk A, Hadjiiski LM, et al. Deep learning in medical imaging and radiation therapy. *Med Phys*. 2019;46:e1–e36.
10. Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. Paper presented at: 2016 Fourth International Conference on 3D Vision (3DV); 2016.

11. Wang S, He K, Nie D, Zhou S, Gao Y, Shen D. CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. *Med Image Anal*. 2019;54:168–178.
12. Philbrick KA, Weston AD, Akkus Z, et al. RIL-contour: a medical imaging dataset annotation tool for and with deep learning. *J Digit Imaging*. 2019;32:571–581.
13. Roth HR, Farag A, Turkbey E, Lu L, Liu J, Summers RM. Data from Pancreas-CT. The cancer imaging archive; 2016.
14. Landman B, Xu Z, Igelsias J, Styner M, Langerak T, Klein A. MICCAI Multi-Atlas Labeling Beyond the Cranial Vault-Workshop and Challenge; 2015.
15. Hu P, Wu F, Peng J, Bao Y, Chen F, Kong D. Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. *Int J Comput Assist Radiol Surg*. 2017;12:399–411.
16. Lessmann N, van Ginneken B, Išgum I.Iterative convolutional neural networks for automatic vertebra identification and segmentation in CT images. Paper presented at: Medical Imaging 2018: Image Processing; 2018.
17. Zheng G, Chu C, Belavý DL, et al. Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: a grand challenge. *Med Image Anal*. 2017;35:327–344.
18. Weston AD, Korfiatis P, Kline TL, et al. Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology*. 2019;290:669–679.
19. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. Paper presented at: International conference on medical image computing and computer-assisted intervention; 2016.
20. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. Paper presented at: International MICCAI Brainlesion Workshop; 2017.
21. Kayalibay B, Jensen G,van der Smagt P. CNN-based segmentation of medical imaging data. arXiv preprint arXiv:170103056; 2017.
22. Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:150500853; 2015.
23. Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:160708022; 2016.
24. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Paper presented at: Proceedings of the thirteenth international conference on artificial intelligence and statistics; 2010.
25. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*. 2004;23:903–921.
26. Kline TL, Korfiatis P, Edwards ME, et al Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys. *J Digit Imaging*. 2017;30:442–448.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig S1. Selected screenshots demonstrating model segmentation performance for the best, median, and worst scans from the test set. Green trace indicates human reference segmentation and red trace indicates model prediction. Values indicate the Dice coefficient.

Table S1. Metrics for all 33 individually segmented organs. IVC, inferior vena cava; SAT, subcutanous adipose tissue; VAT, visceral adipose tissue.